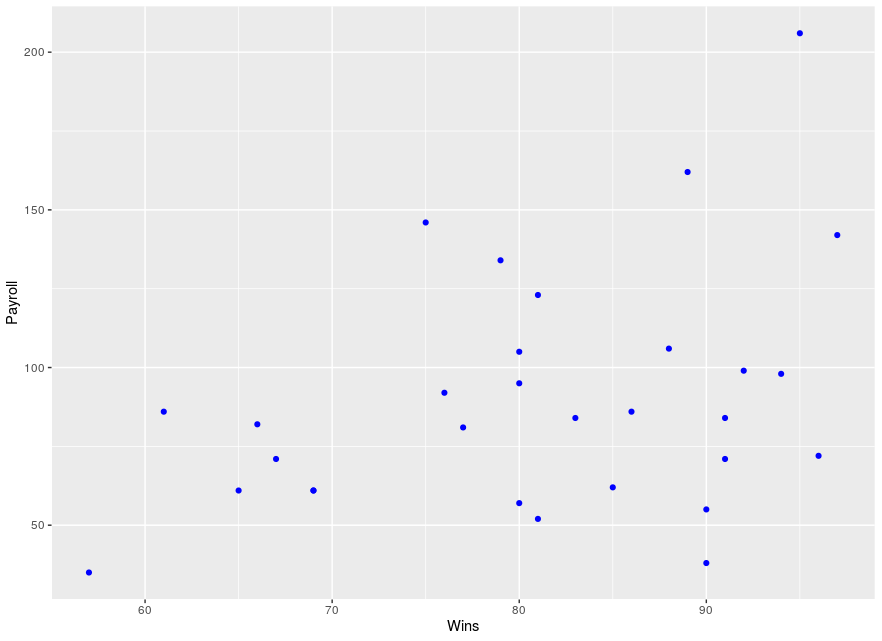
Ben Goodwin

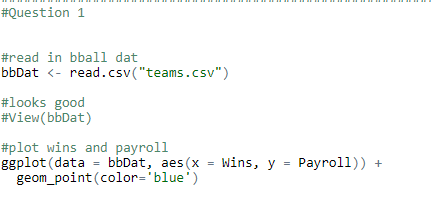
DS6371 HW 8

Question 1)

R output of Payroll and Wins



R input:



SAS input:

data MLB;

input TEAM $ PAYROLL WINS;

datalines;

NYY 206 95

BOS 162 89

CHC 146 75

PHI 142 97

NYM 134 79

DET 123 81

CHW 106 88

LAA 105 80

SF 99 92

MIN 98 94

LAD 95 80

HOU 92 76

SEA 86 61

STL 86 86

ATL 84 91

COL 84 83

BAL 82 66

MIL 81 77

TB 72 96

CIN 71 91

KC 71 67

TOR 62 85

ARZ 61 65

CLE 61 69

WAS 61 69

FA 57 80

TEX 55 90

OAK 52 81

SD 38 90

PIT 35 57

;

run;

PROC sgscatter DATA = MLB;

PLOT PAYROLL \* WINS

/ datalabel = PAYROLL group = WINS;

RUN;

SAS output of Payroll and Wins

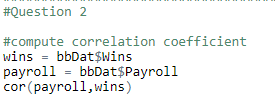
Chart, scatter chart

Description automatically generated

Interpretation: Based on the above plots visualizing the relationship between Wins and payroll, I expect to see positive correlation, although it is very weak, certainly less than 0.5. Why? There is visual evidence of positive correlation, as the data certainly has an observable trend to it. We can see that there appears to be a evidence of a weak linear relationship. I would imagine the correlation coefficient is around 0.2.

Question 2)

R input:



R output:



SAS input:

A picture containing logo

Description automatically generated

SAS output:

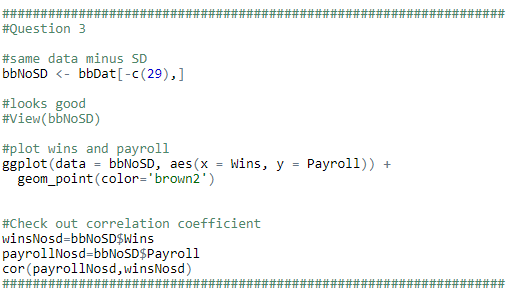
Table

Description automatically generated

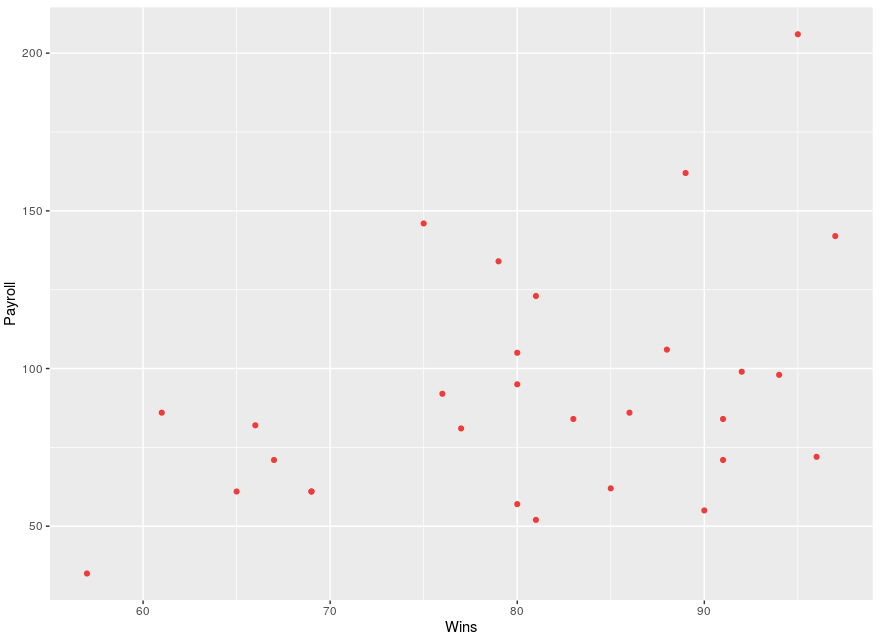
Based on the teams data the correlation coefficient as I calculated is 0.366. This shows that there is positive correlation and that there is evidence of a weak linear relationship.

Question 3)

R input:



R output:





Interpretation: Since we removed one point, the plot did not change significantly. However, removing San Diego from the data changes the correlation coefficient by 0.06. This difference strengthens the correlation in the positive direction.

Question 4)

I think that the league commissioner is taking a single example and trying to generalize it to the broader population of NBA teams. There is also the argument of NYY who had the highest payroll and also the highest number of wins. Aside from Chicago most higher paying teams have more wins. I think more long term sampling is required to determine if this is statement can be backed by fact.

Question 5)

The population for this data are MLB teams. Interestingly there are 30 MLB teams, and the sample here includes all 30 teams. Here we have a rare opportunity where our population is the same as our sample. The population here are MLB teams, along with their wins, and payrolls. These data cannot be considered a random sample since it is the entire population, however, you could use a mechanism to randomly sample this population and draw inference on some other question of interest.